CENTRIST: A Visual Descriptor for Scene Categorization

Jianxin Wu, and James M. Rehg, *Member, IEEE*

J. Wu and J.M. Rehg are with the Georgia Institute of Technology.

Abstract

CENTRIST (CENsus TRansform hISTogram), a new visual descriptor for recognizing topological places or scene categories, is introduced in this paper. We show that place and scene recognition, especially for indoor environments, require its visual descriptor to possess properties that are different from other vision domains (e.g. object recognition). CENTRIST satisfy these properties and suits the place and scene recognition task. It is a holistic representation and has strong generalizability for category recognition. CENTRIST mainly encodes the structural properties within an image and suppresses detailed textural information. Our experiments demonstrate that CENTRIST outperforms the current state-of-the-art in several place and scene recognition datasets, compared with other descriptors such as SIFT and Gist. Besides, it is easy to implement. It has nearly no parameter to tune, and evaluates extremely fast.

Index Terms

Place recognition, scene recognition, visual descriptor, Census Transform, SIFT, Gist

I. INTRODUCTION

Knowing "Where am I" has always being an important research topic in robotics and computer vision. Various research problems have been studied in order to answer different facets of this question. For example, the following three research themes aim at revealing the location of a robot or determining where an image is taken.

- *Place recognition*, or global localization, which identifies the current position and orientation of a robot [1], [2], seeks to find the exact parameterization of a robot's pose in a global reference frame. Place recognition is an inherent part of a Simultaneous Localization and Map Building (SLAM) system [3], [4].
- *Topological place recognition* answers the same question as place recognition, but at a coarser granularity [5]. In topological mapping, a robot is not required to determine its 3D location from the landmarks. It is enough to determine a rough location, e.g. corridor or office 113. A place in topological maps does not necessarily coincide with the human concept of rooms or regions [6]. Topological places are usually generated by a discretization of the robot's environment based on certain distinctive features or events in it.
- *Scene recognition*, or scene categorization, is a term that is usually used to refer to the problem of recognizing the semantic label (e.g. bedroom, mountain, or coast) of a single

image [7], [8], [9], [10], [11]. The input images in scene recognition are usually captured by a person, and are ensured to be representative or characteristic of the underlying scene category. It is usually easy for a person to look at an input image and determine its category label. The learned scene recognizer is generalizable, i.e. it is able to recognize the category of scene images acquired in places that are not present in the training set.

The input in scene recognition are images. Instead, laser range sensors are popular in robot localization tasks. Recently, cameras are also frequently used [12], [13], [14] in robot localization.

In this paper we are interested in recognizing places or scene categories using images taken by usual rectilinear camera lens. Furthermore, since the exact robot pose estimation problem has been widely studied in SLAM systems, we focus on recognizing the topological location or semantic category of a place. Recognizing the semantic category of places from a robot platform is recently emerging as an interesting topic for both vision and robotics research, e.g. visual place categorization [15].

We believe that an appropriate representation (or, more precisely, visual descriptor) is key to the success of a scene recognition task. In the literature, SIFT and Gist are probably the most popular feature descriptors in scene recognition [7], [8], [16], [17], [9], [18], [19], [10], [11], [4]. The SIFT descriptor is originally designed for recognizing the same object appearing under different conditions, and has strong discriminative power. Recognizing topological locations and scene categories, however, poses different requirements for the feature descriptors. Images taken from the same scene category may look very different, i.e. with huge intra-class variations. Similarly, images taken from different part or view point of the same topological location (e.g. office 113) will also contain huge variations. Rather than capturing the detailed textural information of objects in the scene, we would like to capture the the stable spatial structure within images that reflects the functionality of the location [10].

Oliva and Torralba [10] proposed the Gist descriptor to represent such spatial structures. Gist achieved high accuracy in recognizing natural scene categories, e.g. mountain and coast. However, when categories of indoor environments are added, its performance drops dramatically (c.f. Sec. IV-F).

The focus of this paper is CENTRIST (CENsus TRansform hISTogram), a visual descriptor that is suitable for recognizing topological places and scene categories. We will analyze the peculiarity of place images and list a few properties that are desirable for a place/scene recognition representation. We then focus on exhibiting how CENTRIST satisfy these properties better than competing visual descriptors, e.g. SIFT [20], HOG [21] or Gist [10]. We also show that CENTRIST has several important advantages in comparison to state-of-the-art feature descriptors for place/scene recognition and categorization:

- Superior recognition performance on multiple standard datasets;
- Significantly fewer parameters to tune;
- Extremely fast evaluation speed ($>$ 50 fps);
- Very easy to implement, with source code publicly available.

The rest of the paper is organized as follows.[1] Related methods are discussed in Sec. II. Sec. III introduces CENTRIST and focuses on how this visual descriptor suits the place/scene recognition domain. Experiments are shown in Sec. IV. Sec. V concludes this paper with discussions of drawbacks of the proposed method and future research.

II. RELATED WORK

A. *Representation of scene images*

Histograms of various image properties (e.g. color [12], [23], [5], or image derivatives [12]) have been widely used in scene recognition. However, after the SIFT [20] feature and descriptor are popularized in the vision community, it nearly dominates the visual descriptor choice in place and scene recognition systems [7], [8], [17], [9], [18], [19], [11], [4], [24]. SIFT features are invariant to scale and robust to orientation changes. The 128 dimensional SIFT descriptors have high discriminative power, while at the same time are robust to local variations [25]. It has been shown that SIFT descriptor significantly outperforms edge points [9], pixel intensities [7], [8], and steerable pyramids [17] in recognizing places and scenes.

It is suggested in [10] that recognition of scenes could be accomplished by using *global configurations*, without detailed object information. Oliva and Torralba argued for the use of *Shape of the Scene*, an underlying similar and stable spatial structure that presumably exists within scene images coming from the same *function* category, to recognize scene categories. They proposed the Gist descriptor to represent such spatial structures. Gist computed the spectral information in an image through Discrete Fourier Transform (DFT). The spectral signals are then

---

[1] Preliminary versions of portions of this work have been published in [22].

compressed by the Karhunen-Loeve Transform (KLT), a continuous counterpart of the discrete Principal Component Analysis (PCA) method. They showed that many scene signatures such as the degree of *naturalness* and *openness* were reliably estimated from such spectral signals, which in consequence resulted in satisfactory scene recognition results. Since spectral signals were computed from the global image, Oliva and Torralba suggested recognizing scenes without segmentation or recognizing local objects beforehand.

Gist achieved high accuracy in recognizing outdoor scene categories, e.g. mountain and coast. However, when categories of indoor environments are added, the Gist descriptor's performance drops dramatically. We will show in Sec. IV-F that in a 15 class scene recognition dataset [9], which includes the categories used in [10] and several other categories (mainly indoor categories), the accuracy of Gist descriptor is much worse than its performance on outdoor images, and is significantly lower than the proposed CENTRIST descriptor. Our conjecture is that those properties that Gist is modeling are not effective discriminators in indoor environments. For example, almost all indoor images have low degree of naturalness. Similarly, other spatial structure properties modeled by Gist such as the degree of *openness, roughness, and ruggedness* do not apply to indoor scene either.

However, the global configuration argument itself is accepted by many other researchers, whom used the SIFT descriptor to describe the global configuration. Since the SIFT descriptor is designed to recognize the same object instance, statistical analysis of the distribution of SIFT descriptors are popular in scene recognition. Statistics of SIFT descriptors are more tolerant to the huge variations in scene images. SIFT descriptors are first vector quantized to form the *visual codebook* or *visterms*, e.g. by the k-means clustering algorithm. The hope here is that the cluster centers will be meaningful and representative common sub-structures, similar to the codebook in a communication system. We will compare SIFT and CENTRIST in Sec. IV-F.

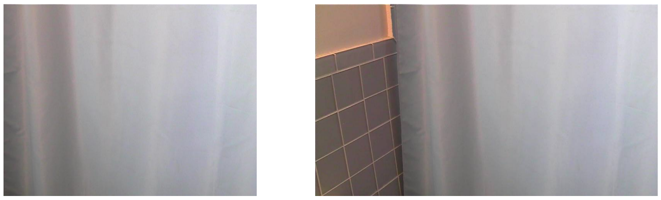
A different representation was proposed by Vogel and Schiele [26]. They split each image into 10 by 10 cells. Each cell was given a semantic label from 9 categories (sky, water, grass, etc.). An SVM classifier ("concept classifier") is then trained to assign labels to new cells. In other words, instead of generating intermediate concepts from data without supervision, they specify a small set of intermediate concepts and learn them in a supervised manner. Category of an image was determined from the concept labels of its 100 cells. Their experiments corroborated the observation that using intermediate concepts gave better performance than using crude image features. However, the concept classifiers' accuracy was lower than 50% in 5 out of the 9 intermediate concepts.

B. Incorporating Spatial Information

SIFT based models usually represent images as *bag of features*, i.e. spatial arrangement information among multiple features are completely ignored. However, it is long recognized that spatial arrangements are essential for recognizing scenes. For example, Szummer and Picard divided images into 4×4 sub-blocks. The K-nearest neighbor classifier was applied to these sub-blocks. The final indoor-outdoor decision was then made based on classification results from the 16 sub-blocks [23]. Their experiments showed that a simple strategy for the second phase classification (majority vote, *i.e.* assigning the image label to the most common class label among the sub-blocks) significantly improved recognition accuracy (approximately 10% higher compared to the sub-block accuracy).

Advocating the global configuration approach, Oliva and Torralba [10] also implicitly used spatial information. In their WDST (Windowed Discriminant Spectral Template, part of the Gist descriptor), spectral information was calculated for 8×8 local patches, with a diameter of 64 pixels for each patch. The Gist descriptor computed from WDST usually outperformed that computed from DST (global Discriminant Spectral Template), sometimes with a large margin. It is natural to conjecture that the spatial arrangement information (implicitly coded in the local WDST ordering) elicited such performance improvements.

Lazebnik, Schmid, and Ponce proposed Spatial Pyramid Matching to systematically incorporate spatial information [9]. Features are quantized into M discrete types using k-means clustering with M centroids. They assume that only features of the same type can be matched. An image is divided in a hierarchical fashion (of level L). The image is divided into $2^l \times 2^l$ sub-blocks in level l, with each dimension (horizontal or vertical) being divided into 2^l evenly sized segments. For a feature type m, X_m and Y_m are sets of the coordinates of type m features. The histogram intersection kernel can be used to compute a matching score for feature type m. The final spatial pyramid matching kernel is then the sum of all such scores. Note that SPM with $L = 0$ reduces to the standard bag of features model.



(a) A partial image (b) The complete image

Fig. 1. A bathroom image is shown in Fig. 1b. Object in the scene (Fig. 1a) does not automatically reveal the room category. This image is best viewed in color.

III. CENTRIST: AN VISUAL DESCRIPTOR FOR PLACE AND SCENE RECOGNITION

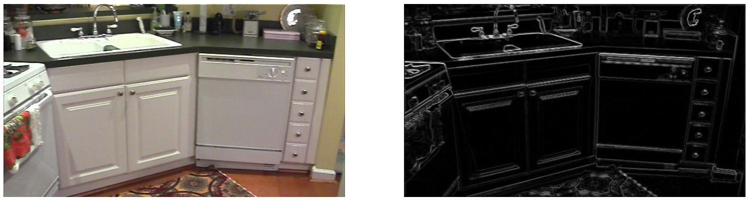
A. Desired properties

We believe that the central problem in place and scene recognition is a visual descriptor that meets the requirements of this domain. In this section, we first discuss some desired properties for a visual descriptor in place and scene recognition tasks. The CENTRIST (CENsus TRansform hISTogram) descriptor is then proposed.

1) Holistic representation: We agree with the "global configurations" approach proposed by Oliva and Torralba [10], which is also followed by quite a few other researchers. [10] recognized outdoor scenes. They found that perceptual properties such as the degree of naturalness can be reliably captured by their holistic representation Gist, and were successfully used to recognize outdoor scene categories. They showed that scene categories can be estimated without explicit recognize objects in the scene. These perceptual properties are not as useful in indoor environments. For example, indoor environments all share low degree of naturalness. However, we believe that a holistic approach should still be used.

As illustrated in Fig. 1, knowing the object in an image does not automatically tell us the place category. Instead, the curtain object and the tiles on the wall altogether clearly show that this is a bathroom image. Many useful information sources such as the tiles are usually contained in those regions that are not objects. Furthermore, recognizing objects in cluttered environments is not necessarily easier than scene recognition itself. Thus we prefer a holistic representation.

2) Capturing the structural properties: One property of the place categorization problem is that huge variations exist within the same place category. As a consequence, we can not search for a template pattern using the approaches applied in most object recognition tasks [27]. For example, a bed will likely appear in an image taken from a bedroom. However, the picture can be taken from any viewpoint, and the bed can appear in many variations (mattress, sofa bed,



(a) An example kitchen image (b) Corresponding Sobel image

Fig. 2. Fig. 2a shows an example kitchen image. Fig. 2b shows its corresponding Sobel gradients. The Sobel gradients are normalized to $[0 \quad 255]$. This image is best viewed in color.

water bed mattress, etc.). Bedrooms can be decorated in any possible color or style, with very different furniture and illumination conditions. These variations pose difficulties for the local patch based representations (e.g. SIFT).

To make the problem even more difficult, since we are interested in autonomous data collection, the images may not be representative of the place category. In the bedroom example, a bed could possibly be invisible in many frames. These difficulties suggest that we need a visual descriptor that suppresses fine scale textural details and focuses on structural properties. For example, beds provide the same function to people. Although they may show large visual differences, they are mostly rectangular, with a relatively flat surface, and having pillows. These common properties exist because they are necessary for providing a bed's functionality. In turn, these properties exhibit similar structural properties in bed images.

In other words, we want the descriptor to (implicitly or explicitly) capture more general structural properties such as rectangular shapes, flat surfaces, tiles, etc., while suppressing detailed textural information. In recognizing place categories, these fine-scaled textures will distract the classifier. They can be noisy and harmful if the feature extraction method is not carefully designed. Fig. 2 further illustrates this idea. As shown in Fig. 2b, the spatial structure is more prominent in the Sobel image, e.g. the shape that reflects the sink and dishwasher. It is possible to recognize the kitchen category from the Sobel image alone.

It is worth noting that most of the perceptual properties used for scene recognition in [10] are well preserved in Sobel images too. For example, the *degree of naturalness*, defined by the distribution of edges, was used to recognize scenes in [10]. In comparison to the original images, it is easy to read out from the Sobel images that man-made environments have more horizontal and vertical edges, thus they have lower degree of naturalness. Similarly, other spatial structure properties such as the degree of *openness, roughness, and ruggedness* are also easy to capture

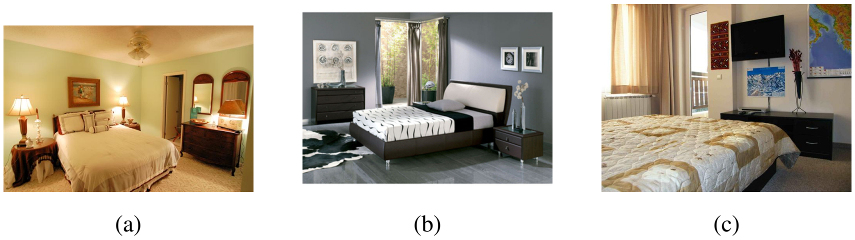

(a) (b) (c)

Fig. 3. Place images do not exhibit strong geometry constraints among objects.

in Sobel images.

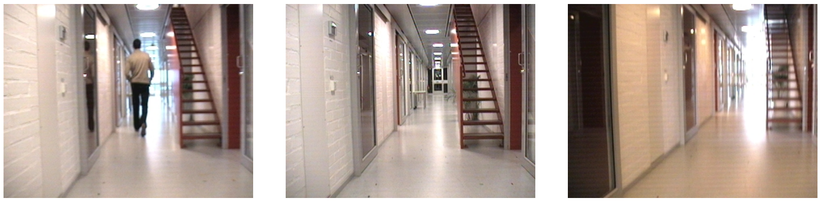
We are, however, not proposing to use Sobel images directly as a descriptor. On the one hand, structural properties (mainly boundaries/edges) are crucial for recognizing place categories. On the other hand, we need a better descriptor that summarizes such information efficiently. In Fig. 2b we observe that such structural properties can be reflected by the distribution of local structures. For example, the percentages of local structures that are local horizontal edge, vertical edge, or junctions. Our CENTRIST descriptor will model the distribution of local structures.

3) Rough geometry is useful: Strong geometrical constraints (e.g. the constellation model [28] or pictorial structures [29]) are very useful in object recognition. However, they are essentially not applicable in place categorization due to the large variations. Different objects can be arranged in any spatial configuration in a place image. As shown in Fig. 3, the furniture and objects in a bedroom can be arranged in arbitrary configurations.

However, rough geometry constraints are very helpful in recognizing place categories [9]. For example, a reading lamp is usually placed close to the bed in a bedroom as in Fig. 3a. While a TV appears in a bedroom, it is often at the foot of the bed as shown in Fig. 3c. More general constraints such as *The sky should be on top of the ground* will help reduce ambiguity, even when the images are taken from random viewpoints;

4) Generalizability: The learned category concepts will be applied to new images. An ideal situation is that the feature descriptors are compact within a category (even under large visual variations), and are far apart when they belong to different categories.

Fig. 4 shows three images from the corridor environment. These images are taken from approximately the same position in the same environment, and we already see large visual variations. We would expect even larger variations from pictures taken in different corridor environments. However, the visual descriptor must be able to capture these similar spatial structures: open spaces in the middle, stairs, strips on the wall, etc.



(a) Cloudy (b) Night (c) Sunny

Fig. 4. Example images from the KTH IDOL dataset. Images showed approximately the same location under different weather conditions. Images were taken by a robot called Minnie.

We propose to use CENTRIST (CENsus TRansform hISTogram) as our visual descriptor for the place category recognition task. CENTRIST is a holistic representation that captures structural properties by modeling distribution of local structures. We capture rough geometrical information by using a spatial CENTRIST representation. CENTRIST also has similar descriptor vectors for images in the same place category. We will now introduce CENTRIST in the following sections, and show how CENTRIST satisfy the desired properties we have just discussed.

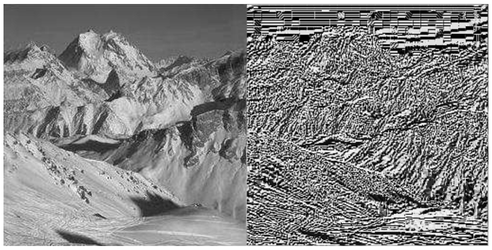
B. Census Transform (CT) and CENTRIST

Census Transform (CT) is a non-parametric local transform originally designed for establishing correspondence between local patches [30]. Census transform compares the intensity value of a pixel with its eight neighboring pixels, as illustrated in Eqn. 1. If the center pixel is bigger than (or equal to) one of its neighbors, a bit 1 is set in the corresponding location. Otherwise a bit 0 is set.

$$\begin{array}{c|c|c} 32 & 64 & 96 \\ \hline 32 & \mathbf{64} & 96 \\ \hline 32 & 32 & 96 \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array} \Rightarrow (11010110)_2 \Rightarrow \text{CT} = 214 \tag{1}$$

The eight bits generated from intensity comparisons can be put together in any order (we collect bits from left to right, and from top to bottom), which is consequently converted to a base-10 number in $[0\ 255]$. This is value is the Census Transform value (CT value) for this center pixel. Similar to other non-parametric local transforms which are based on intensity comparisons (e.g. ordinal measures [31]), Census Transform is robust to illumination changes, gamma variations, etc. Note that the Census Transform is equivalent (modulo a difference in bit ordering) to the *local binary pattern* code $LBP_{8,1}$ [32].

As a visualization method, we create a *Census Transformed image* by replacing a pixel with its CT value. Shown by the example in Fig. 5, the Census Transform retains global structures



(a) mountain image (b) Transformed image

Fig. 5. An example *Census Transformed image*.

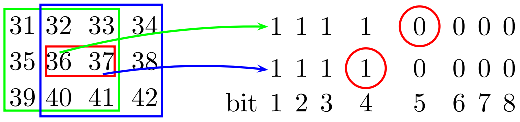


Fig. 6. Illustration of constraints between CT values of neighboring pixels. (This picture is best viewed in color.)

of the picture (especially discontinuities) besides capturing the local structures as it is designed for.

Another important property of the transform is that CT values of neighboring pixels are highly correlated. In the example of Fig. 6, we examine the direct constraint posed by the two center pixels. The Census Transform for pixels valued 36 and 37 are depicted in right, and the two circled bits are both comparing the two center pixels (but in different orders). Thus the two circled bits are constrained to be strictly complement to each other if the two pixels are not equal. More generally, bit 5 of CT(x, y) and bit 4 of CT$(x + 1, y)$ must always be complementary to each other, since they both compare the pixels at (x, y) and $(x + 1, y)$ if these two pixels are not equal. There exist many other such constraints. In fact, there are eight such constraints between one pixel and its eight neighboring pixels.

Besides these deterministic constraints, there also exist indirect constraints that are more complex. For example, in Fig. 6, the pixel valued 32 compares with both center pixels when computing their CT values (bit 2 of CT(x, y) and bit 1 of CT$(x + 1, y)$). Depending on the comparison results between the center pixels, there are probabilistic relationships between these bits.

The transitive property of such constraints also make them propagate to pixels that are far apart. For example, in Fig. 6, the pixels valued 31 and 42 can be compared using various paths of comparisons, e.g. $31 < 35 < 39 < 40 < 41 < 42$. Similarly, although no deterministic comparisons can be deduced between some pixels (e.g. 34 and 39), probabilistic relationships still can be obtained. The propagated constraints make Census Transform values and Census Transform histograms implicitly contain information for describing global structures, unlike the histogram of pixel values.

Finally, the Census Transform operation transforms any 3 by 3 image region into one of 256 cases, each corresponding to a special type of local structure of pixel intensities. The CT value acts as an index to these different local structures. No total ordering or partial ordering exists among the CT values. It is important to refrain from comparing two CT values as comparing two integers (like what we do when comparing two pixel intensity values).

A histogram of CT values for an image or image patch[2] can be easily computed, and we use CENTRIST (CENsus TRansform hISTogram) as our visual descriptor. CENTRIST can be computed very efficiently. It only involves 16 operations to compute the CT value for a center pixel (8 comparisons and 8 additional operations to set bits to 0 or 1). The cost to compute CENTRIST is linear in the number of pixels of the region we are interested in. There is also potential for further acceleration to the computation of CENTRIST, by using special hardware (e.g. FPGA), because it mainly involves integer arithmetic that are highly parallel.

C. Constraints among CENTRIST components

Usually there is no obvious constraint among the components of a histogram. For example, we would often treat the R, G, and B components of a color histogram independent to each other. CENTRIST, however, exhibits strong constraints or dependencies among its components.

Take as example the direct constraint shown in Fig. 6, bit 5 of CT(x, y) and bit 4 of CT$(x, y{+}1)$ must be complementary to each other if they are not equal. Both bits are 1 if they are equal. If we apply this constraint to all pixels in an image, we get to the conclusion that *the number of pixels whose CT value's bit 5 is 1 must be equal to or greater than*[3] *the number of pixels whose CT value's bit 4 is 0*, if we ignore border pixels where such constraints break. Let $\boldsymbol{h}$ be the CENTRIST descriptor of any image. The above statement is translated into the following

---

[2] In fact, it can be computed for an image region of arbitrary shape.

[3] These extra 1's are caused by the special case when two neighboring pixels are equal to each other.

equation:

$$\sum_{i \text{ \& 0x08 = 0x08}} \boldsymbol{h}(i) \ \geq \sum_{i \text{ \& 0x10 = 0}} \boldsymbol{h}(i), \tag{2}$$

where $\&$ is *bitwise and*, 0x08 is the number 08 in hexadecimal format, and $0 \leq i \leq 255$. Thus the left hand side of Eqn. 2 counts the number of pixels whose CT value's bit 5 is 1. By switching 1 and 0, we get another equation:

$$\sum_{i \text{ \& 0x08 = 0}} \boldsymbol{h}(i) \ \leq \sum_{i \text{ \& 0x10 = 0x10}} \boldsymbol{h}(i). \tag{3}$$

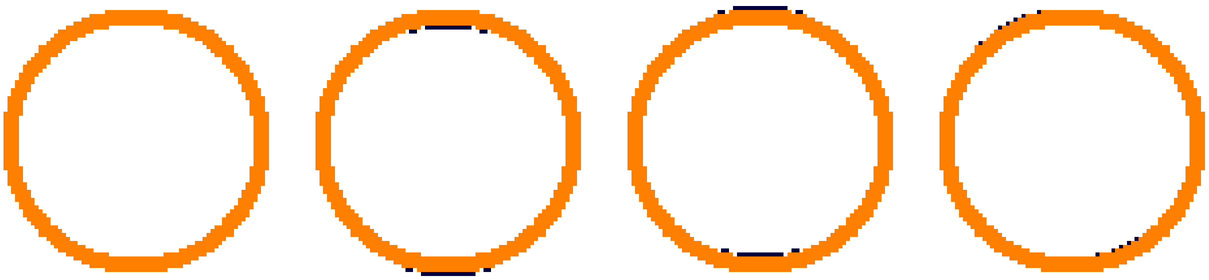
Similarly, 6 other linear inequalities can be specified by comparing CT(x, y) with CT$(x-1, y-1)$, CT$(x-1, y)$, and CT$(x-1, y+1)$. Thus, any CENTRIST feature vector resides in a subspace that is defined by these linear inequalities.

We can not write down explicit equations for the indirect or transitive constraints in a CENTRIST feature vector. However, we expect these constraints will further reduce the dimension of the subspace of CENTRIST feature vectors. A CENTRIST feature vector, although having 256 bins, is living in a subspace whose dimension is much lower than 256. The constraints among elements in a CENTRIST make PCA suitable for dimension reduction.

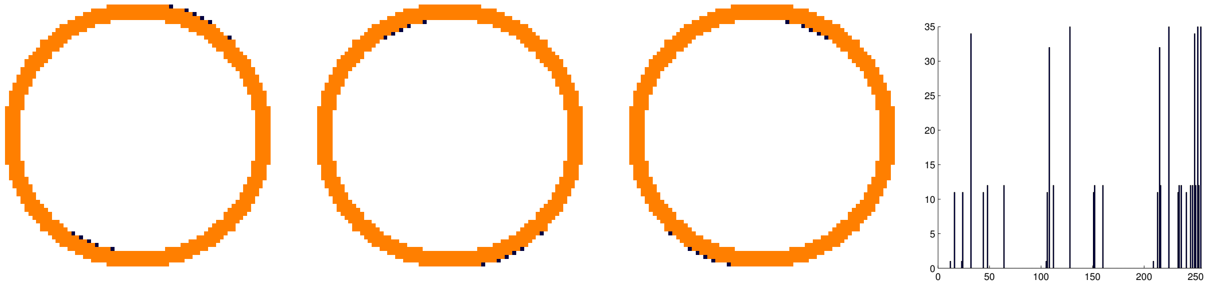
D. CENTRIST encodes image structures

In order to understand why CENTRIST efficiently captures the essence of a scene image, it is worthwhile to further examine the distribution of CT values and CENTRIST feature vectors. Using images from the 15 class scene dataset [9], we find that the 6 CT values with highest frequencies are CT $= 31, 248, 240, 232, 15, 23$ (excluding 0 and 255). As shown in Fig. 7b-7g, these CT values correspond to local 3×3 neighborhoods that have either horizontal or various close-to-diagonal edge structures. It is counter-intuitive that vertical edge structures are not among the top candidates. A possible explanation is that vertical structures are usually appearing to be inclined in pictures because of the perspective nature of cameras.

CENTRIST of the example ellipse image (Fig. 7a) is shown in Fig. 7h. It summarizes the distribution of various local structures in the image. Because of the strong correlation of neighboring CT values, the histogram cells are not independent of each other. A CENTRIST feature vector implicitly encodes strong constraints of the global structure of the image. For example, if an image has a CENTRIST feature vector close to that of Fig. 7h, we would well

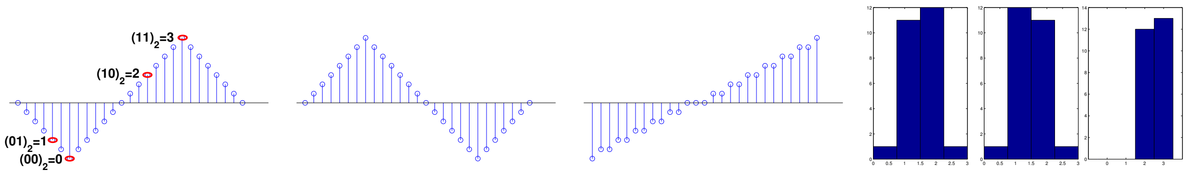


(a) ellipse (b) CT = 31 (c) CT = 248 (d) CT = 240



(e) CT = 232 (f) CT = 15 (g) CT = 23 (h) CENTRIST

Fig. 7. Illustration of Census transforms. Fig. 7a is an example image of ellipse. Fig. 7b-7g show pixels having the 6 highest frequency CT values (shown in black). Fig. 7h is the CENTRIST feature vector of Fig. 7a. (This image is best viewed in color.)



(a) (b) (c) (d) CENTRIST

Fig. 8. Census Transform encodes shape in 1-d. The Census Transform values of (a)-(c) are shown in the caption. Sub-figure (d) shows CENTRIST descriptors of figures (a)-(c), respectively. Both end points are ignored in compute CT. (This image is best viewed in color.)

expect the image to exhibit an ellipse shape with a high probability (c.f. Sec. III-E for more evidences.)

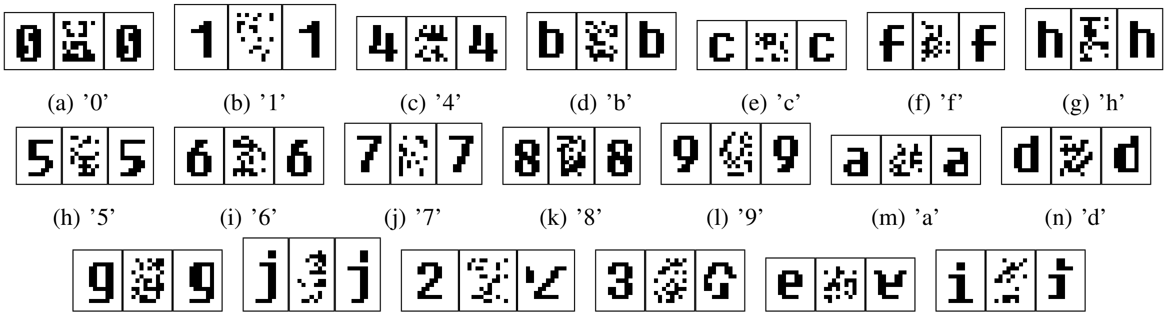
A simplification to the one dimensional world better explains the intuition behind our statement. In 1-d there are only 4 possible CT values, and the semantic interpretation of these CT values are obvious. As shown in Fig. 8a, the four CT values are $CT = 0$ (valley), $CT = 1$ (downhill), $CT = 2$ (uphill), and $CT = 3$ (peak). For simple shapes in 1-d, CENTRIST encodes shape information and constraints. Downhill shapes and uphill shapes can only be connected by a valley, and uphill shapes require a peak to transit to downhill shapes. Because of these constraints, the only other shapes that have the same CENTRIST descriptor as that of Fig. 8a are those shapes that cut a small portion of the left part of Fig. 8a and move it to the right. Images that are different but keep the shapes (e.g. Fig. 8b) also are similar in their CENTRIST descriptors (Fig. 8d). On the contrary, a huge number of possible curves have the same intensity histogram as that of Fig. 8a. Even if we impose smoothness constraints between neighboring pixel intensities, the shape ambiguity is still large (e.g. Fig. 8c is smooth and has the same intensity histogram as that of Figs. 8a and 8b, but it has different shape and a very different CENTRIST descriptor).

E. Reconstructing image patches from CENTRIST descriptors

It is well known that spatial information is totally lost in the histogram of pixel intensities. CENTRIST, however, implicitly retains the global spatial structure of an image patch through the constraints we have discussed. We performed some reconstruction experiments to further illustrate this idea. When we randomly shuffle the pixels of an input image, the original structure of the image is completely lost. Using the shuffled image as an initial state, we repeatedly change two pixels at one time, until the current state has the same CENTRIST descriptor as the input image. This optimization is guided by the Simulated Annealing algorithm, and the algorithm terminates when the current state has the same CENTRIST descriptor as the input image. If structure of the original image is observed in the reconstruction result (i.e. the termination state), this is an evidence that structure of an image is (at least partially) encoded in its CENTRIST descriptor.

In the reconstruction results in Fig. 9, the left image in each subfigure is the input image. A pair of pixels in the input images are randomly chosen and exchanged. The exchange operation is repeated multiple times (equal to the number of pixels in the input image), which gives the initial state for the reconstruction. Our goal is to find an image that has the same CENTRIST descriptor as the input. Thus the cost function is set to the Euclidean distance between CENTRIST descriptors of the current state and the input. The terminating state is output of the reconstruction (right image in each subfigure of Fig. 9). Note that our reconstruction experiments are analogous to the histogram matching approach for texture synthesis of Heeger and Bergen [33].

Although the initial states look like random collection of pixels, many of the reconstruction results perfectly match the input images subfigure (a)-(g) in Fig. 9). More examples are recon-


(a) ’0’ (b) ’1’ (c) ’4’ (d) ’b’ (e) ’c’ (f) ’f’ (g) ’h’

(h) ’5’ (i) ’6’ (j) ’7’ (k) ’8’ (l) ’9’ (m) ’a’ (n) ’d’

(o) ’g’ (p) ’j’ (q) ’2’ (r) ’3’ (s) ’e’ (t) ’i’

Fig. 9. Reconstruct images from CENTRIST descriptors. In each group of images, the left image is the input image. The image in the middle is the initial starting point that is generated by randomly exchanging pairs of pixels in the input. The reconstruction result is shown in the right of each group, which has the same CENTRIST descriptor as the input image.

structed with minor discrepancies (subfigure (h)-(p) in Fig. 9). Large scale structures of the input digits and characters are successfully reconstructed in these images, with small errors. In the rest examples, e.g. ‘2’ and ‘e’, major structures of the original input images are still partially revealed. These results empirically validated that CENTRIST descriptors have the ability to encode the shape of an image.

An analogy to these results is the jigsaw puzzle. The CT value in each pixel location is analogous to a puzzle piece of certain type. Pieces can be put next to each other only if their shapes satisfy certain constraints (similar to the constraints between neighboring CT values). After breaking a puzzle into pieces, there are only a very limited number of ways to assemble the pieces together, and we would expect the assembled version to resemble the original one with a high probability.

Similarly, there are a huge number of ways to shuffle pixels of an input image (possibly exponential in the number of pixels). However, if we add an additional constraint that the CENTRIST descriptor should be same as the input image, there is only a small number of possibilities. For example, at least 2 images have the same CENTRIST descriptor as the digit ‘9’ (the original input digit ‘9’ and the slightly different reconstruction). As shown in Fig. 9, these remaining reconstructions have a large chance to share same or similar structure as the input image.

Two points are worth pointing out about the reconstruction results. First, in larger images a CENTRIST descriptor is not enough to reconstruct the original image.[4] However, as a feature

---

[4] Note that the images in Fig. 9 are black-and-white images instead of gray-scale ones.

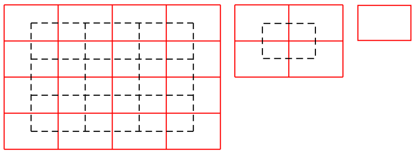


Fig. 10. Illustration of the level 2, 1, and 0 split of an image.

descriptor, it has the ability to distinguish between images with different structural properties. Second, it is essentially impossible to reconstruct even a small image using other descriptors (e.g. SIFT, HOG [21], or Gist).

F. Spatial representations

Because CENTRIST can only encode global shape structure in a small image patch, in order to capture the global structure of an image in larger scale, we propose a spatial representation based on the Spatial Pyramid Matching scheme [9]. A *spatial pyramid* (dividing an image into subregions and integrating correspondence results in these regions) encodes rough global structure of an image and usually improves recognition. It is straightforward to build a spatial pyramid for the proposed CENTRIST representation.

As shown in Fig. 10, the level 2 split in a spatial pyramid divides the image into $2^2 \times 2^2 = 16$ blocks. We also shift the division (dash line blocks) in order to avoid artifacts created by the non-overlapping division, which makes a total of 25 blocks in level 2, which is different from the spatial hierarchy proposed in [9]. Similarly, level 1 and 0 have 5 and 1 blocks respectively. The image is resized between different levels so that all blocks contain the same number of pixels. CENTRIST in all blocks are then concatenated to form an overall feature vector. For example, we use PCA to reduce the dimensionality of CENTRIST to 40, a level 2 pyramid will then result in a feature vector which has $40 \times (25 + 5 + 1) = 1240$ dimensions.

We want to emphasize that this spatial representation is independent of the descriptor used for each sub-window. In this paper, we use two different representations. In the first we use PCA to reduce the CENTRIST descriptor to 40 dimensions, which we call PACT (Principal component Analysis of Census Transform histograms). In the second approach we use a bag of visual words model with CENTRIST as the base visual descriptor. Details of both approaches are described in Sec. IV. Note that other visual descriptors can also be used with the spatial representation.

G. Limitations of CENTRIST

As we have stated from the very beginning, CENTRIST is designed to be a representation that suits place recognition and categorization problems. This design choice renders limitations that prevent it from being applied in some applications. We list the limitations below, and explain how these limitations affect the place and scene recognition performance.

- CENTRIST is sensitive to rotations. Thus it is not suitable for 3-D or multiview object recognition, e.g. the Caltech 101 dataset [34]. In scene recognition, images are always taken in the upright view and we usually pay attention to the overall structure of the scene, which is not prone to rotational variances. Furthermore, CENTRIST is invariant to translation and robust against scale changes;
- CENTRIST is not a precise shape descriptor. It is designed to recognize shape categories (*c.f.* Sec. IV-A), but not for exact shape registration applications, *e.g.* the shape retrieval task in [35], [36].
- CENTRIST ignores color information.

IV. EXPERIMENTS

The CENTRIST visual descriptor is tested on 4 datasets: Swedish leaf [37], KTH IDOL [12], 15 class scene category [9], and the 8 class sports event dataset [18]. In each dataset, the available data are randomly split into a training set and a testing set following published protocols on these datasets. The random splitting is repeated 5 times, and the average accuracy is reported. Although color images are available in 3 datasets (leaf, IDOL, and events), we only use the intensity values and ignore color information.

Our first approach to apply CENTRIST uses PCA (Principal Component Analysis) to reduce its dimensionality to 40. In computing CENTRIST descriptors and PCA eigenvectors, we remove two bins with CT $= 0, 255$ and normalize the CENTRIST descriptors and PCA eigenvectors such that they have zero mean and unit norm. Also, we do not subtract mean in PCA for computational efficiency. Our experiments show that this does not cause significant difference in recognition results.[5] CENTRIST will also be used in a Bag of Visual words framework in Sec. IV-E.

---

[5] PACT Code is available at `http://www.cc.gatech.edu/~wujx/PACT/PACT.htm`. Please refer to the code for details.

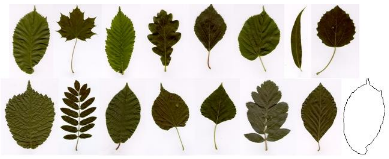


Fig. 11. Example images from the Swedish Leaf dataset. The first 15 images are chosen from the 15 leaf species, one per species. The last image is the contour of the first leaf image.

After the spatial PACT feature vectors are extracted from images, we choose different classifiers for recognizing place instances and categories, in order to find the right tradeoff between discriminative power and invariance for both problems. In recognizing topological places, we use the Nearest Neighbor classifier (1-NN, to be precise). Thus we are looking for places that have not only similar local patches, but also *exact* spatial arrangements of these patches. SVM classifiers are used for scene category recognition. In category recognition we are only expecting loose spatial information. We rely on the generalization ability of SVM to capture such relationships, while avoiding overfitting.

Since the CT values are based solely on pixel intensity comparisons, it might be helpful to include a few image statistics, e.g. average value and standard deviation of pixels in a block. We append these statistics to spatial PACT in the input to SVM classifiers for scene recognition problem. The feature vector of a level 2 spatial PACT then becomes $(40+2)\times(25+5+1)=1302$ dimensional. However, both image statistics are not used in the 1-NN classifiers, since the large variation of illumination in place instance recognition tasks will cause these global statistics to be unreliable.

A. Swedish Leaf

The Swedish leaf dataset [37] collects pictures of 15 species of Swedish leaves (c.f. Fig. 11). There are 75 images in each class. Following the protocol of [37], 25 images from each class are used for training and the rest 50 for testing. This dataset has been used to evaluate shape matching methods [35], [36], in which the contour of leaves (instead of the gray-scale or color leaf picture) were used as input (e.g. the last picture in Fig. 11). In the contour image, no other information is available (e.g. color, texture) besides shape or structure of the leaf. We use the contour input to further verify our statement that the visual descriptor encodes such information.

The first 25 images from each class are used to train the PCA eigenvectors. 10 and 40

TABLE I

RESULTS ON THE SWEDISH LEAF DATASET.

Method	Input	Rates
Shape-Tree [35]	Contour only	**96.28%**
IDSC+DP [36]	Contour only	94.13%
spatial PACT	Contour only	90.77%
SC+DP [36]	Contour only	88.12%
Söderkvist [37]	Contour only	82.40%
spatial PACT	Gray-scale image	**97.92%**
SPTC+DP [36]	Gray-scale image	95.33%

eigenvectors are used when the inputs are contour and intensity images, respectively. Results on this dataset are shown in Table I. Although not specifically designed for matching shapes, spatial PACT can achieve 90.77% accuracy on leaf contours, better than Shape Context+Dynamic Programming (SC+DP). When pictures instead of contours are used as input, spatial PACT can recognize 97.92% leaves, which outperforms other methods by a large margin.

B. KTH IDOL and INDECS

The KTH IDOL (Image Database for rObot Localization) dataset [38] was captured in a five-room office environment, including a one-person office, a two-person office, a kitchen, a corridor, and a printer area. Images were taken by two Robots: Minnie and Dumbo. The purpose of this dataset is to recognize which room the robot is in based on a single image, i.e. a topological place instance recognition problem.

Cameras were mounted at different heights on the robots, which made the pictures taken by the two robots quite different. Image resolution was 320×240. A complete image sequence contained all the images captured by a robot when it was driven through all five rooms. Images were taken under 3 weather conditions: Cloudy, Night, and Sunny. For each robot and each weather condition, 4 runs of robot driving were taken on different days. Thus, there are in total $2 \times 3 \times 4 = 24$ image sequences. Various changes during different robot runs (e.g. moving persons, changing weather and illumination conditions, relocated/added/removed furniture make this dataset both realistic and challenging. Fig. 4 in page 10 shows images taken by the Minnie robot under 3 different weather conditions at approximately the same location, but with substantial visual

TABLE II

AVERAGE ACCURACIES ON RECOGNIZING PLACE INSTANCES USING THE KTH-IDOL DATASET AND THE KTH-INDECS DATASET. LEVEL 2 PYRAMIDS ARE USED FOR SPATIAL PACT. "ROBOTS" MEANS BOTH MINNIE AND DUMBO.

Train	Test	Condition	spatial PACT+1-NN	spatial PACT+SVM	[12]
Minnie	Minnie	Same	95.35%	94.79%	**95.51%**
Dumbo	Dumbo	Same	**97.62%**	96.35%	97.26%
Minnie	Minnie	Different	**90.17%**	83.10%	71.90%
Dumbo	Dumbo	Different	**94.98%**	89.35%	80.55%
Minnie	Dumbo	Same	**77.78%**	70.15%	66.63%
Dumbo	Minnie	Same	**72.44%**	65.18%	62.20%
Camera	Camera	Different	**90.01%**	78.39%	75.67%
Camera	Robots	Same	**64.39%**	42.16%	50.56%

changes.

In our experiments we use the run 1 and 2 in each robot and weather condition. We perform 3 types of experiments as those in [12]. First we train and test using the same robot, same weather condition. Run 1 is used for training and run 2 for testing, and vice versa. Second we use the same robot for training and testing, but with different weather conditions. These experiments test the ability of spatial PACT to generalize over variations caused by person, furniture, and illumination. The third type of experiment uses training and testing set under the same weather conditions, but captured by different robots. Note that images taken by the two robots are quite different.

The KTH-INDECS dataset [39] was collected in the same environment with IDOL. Instead of using robots, cameras were mounted in several fixed locations inside each room. Pictures of multiple viewing angles were taken in each location. In the last type of experiment we use INDECS images as training examples, and test on both INDECS images under different weather conditions and on images taken by robots. Results using level 2 pyramid spatial PACT and 1-NN are shown in Table II, compared against results in [12].

In the first type of experiments, both spatial PACT and the method in [12] attain high accuracy ($> 95\%$), and the two methods are performing roughly equally well. However, in the second type of experiments spatial PACT has significantly higher accuracies (18% higher in Minnie and 14% higher in Dumbo). The superior performance of our CENTRIST based representation shows

TABLE III

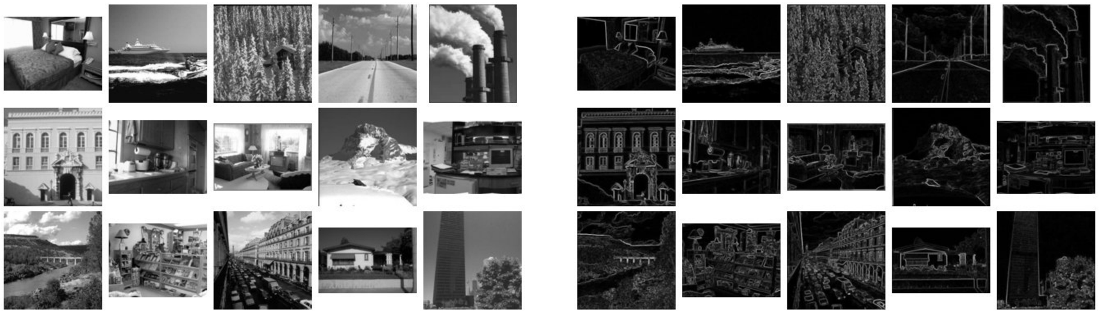
AVERAGE ACCURACIES ON THE KTH-IDOL DATASET AND THE KTH-INDECS DATASET USING DIFFERENT LEVELS OF SPATIAL PYRAMID. "ROBOTS" MEANS BOTH MINNIE AND DUMBO.

Train	Test	Condition	$L = 3$	$L = 2$	$L = 1$	$L = 0$
Minnie	Minnie	Same	95.01%	**95.35%**	95.08%	86.08%
Dumbo	Dumbo	Same	95.51%	**97.62%**	96.87%	88.26%
Minnie	Minnie	Different	**90.30%**	90.17%	85.75%	60.51%
Dumbo	Dumbo	Different	94.67%	**94.98%**	91.75%	74.67%
Minnie	Dumbo	Same	74.96%	**77.78%**	75.56%	62.34%
Dumbo	Minnie	Same	68.59%	**72.44%**	71.36%	53.74%
Camera	Camera	Different	**92.37%**	90.01%	84.80%	71.45%
Camera	Robots	Same	60.73%	**64.39%**	57.87%	41.55%

that it is robust to illumination changes and other minor variations (e.g. moving persons, moved objects in an image, etc). The Dumbo robot achieves a 94.57% accuracy using a single input image without knowing any image histories (a "kidnapped robot" [13]). Thus, after walking a robot in an environment, spatial PACT enables the robot to robustly answer the question "Whare am I?" based on a single image, a capacity that is very attractive to indoor robot applications. When the training and testing data come from different robots, the performance of both methods drop significantly. This is expected, since the camera heights are quite different. However, spatial PACT still outperforms the SVM classifier in [12] by about 10%. In the last type of experiment involving camera images, spatial PACT achieved about 14% higher accuracies than those reported in [12].

We also tested the effects of using different pyramid levels. As shown in Table III, applying a spatial pyramid matching scheme greatly improves system performances ($L > 0$ vs. $L = 0$). However, the improvement after $L > 2$ is negligible. $L = 3$ performance is even worse than that of $L = 2$ in most cases. Our observation corroborates that of Lazebnik, Schmid and Ponce in [9], which used a scene recognition dataset. In the remainder of this paper, we will use $L = 2$ in spatial PACT.

CENTRIST can be computed and evaluated quickly, and so is spatial PACT. The IDOL dataset has around 1000 images in each image sequence, and spatial PACT processes at about 50 frames per second on an Intel Pentium 4 2GHz computer for computing the features, and finding the



(a) images from 15 scene categories (b) corresponding Sobel images

Fig. 12. Images from 15 different scene categories. Fig. 12a shows one image from each of the 15 categories from [9]. The categories are bedroom, coast, forest, highway, industrial, inside city, kitchen, living room, mountain, office, open country, store, street, suburb, and tall building, respectively (from top to bottom, and from left to right). Fig. 12b shows their corresponding Sobel gradient images. The Sobel gradients are normalized to $[0 \ \ 255]$.

1-NN match.[6]

Finally, we have discussed in Sec. III-F that 1-NN is chosen for the KTH IDOL dataset because we want to match the global geometric structure among image regions in addition to local image patches. Table II confirms that in this dataset the 1-NN classifier outperforms SVM with a large margin.

C. The 15 class scene category dataset

The 15 class scene recognition dataset was built gradually by Oliva and Torralba ([10], 8 classes), Fei-Fei and Perona ([8], 13 classes), and Lazebnik, Schmid and Ponce ([9], 15 classes). This is a scene category dataset (scene classes including office, store, coast, etc. Please refer to Fig. 12 for example images and category names.) Images are about 300×250 in resolution, with 210 to 410 images in each category. This dataset contains a wide range of scene categories in both indoor and outdoor environments. Unlike the KTH IDOL images which are taken by robots, images in this datasets are taken by people and representative of the scene category. We use SVM and spatial PACT in this dataset. The first 100 images in each category were used to perform PCA. Same as previous research on this dataset, 100 images in each category are used for training, and the remaining images constitute the testing set. The results are shown in

---

[6] Or 20 fps if including the time for loading the test image from hard drive.

TABLE IV

RECOGNITION RATES ON THE 15 CLASS SCENE DATASET.

L	Method	Feature type	Rates
0	SPM [9]	16 channel weak features	45.3 ± 0.5
0	SPM [9]	SIFT, 200 cluster centers	72.2 ± 0.6
0	SPM [9]	SIFT, 400 cluster centers	$\mathbf{74.8 \pm 0.3}$
0	spatial PACT	CENTRIST, 40 eigenvectors	73.29 ± 0.96
3	SPM [9]	16 channel weak features	66.8 ± 0.6
2	SPM [9]	SIFT, 200 cluster centers	81.1 ± 0.3
2	SPM [9]	SIFT, 400 cluster centers	81.4 ± 0.5
3	SPM [19]	SIFT, 400 intermediate concepts	83.3
2	spatial PACT	CENTRIST, 40 eigenvectors	83.10 ± 0.60
2	SP-pLSA [7]	SIFT, 1200 pLSA topics	83.7
2	spatial PACT with Sobel image	CENTRIST, 40 eigenvectors	$\mathbf{84.96 \pm 0.34}$

Table IV, where our level 2 pyramid spatial PACT achieves the highest accuracy[7].

In [9], low level features were divided into weak features (computed from local 3×3 neighborhoods) and strong features (SIFT descriptors computed from 16×16 image patches). Strong features were shown to have much higher accuracy than weak features (c.f. Table IV). The Census Transform is computed from 3×3 local neighborhoods, and falls into the weak feature category. However, when $L = 0$ (not using spatial pyramid), CENTRIST substantially outperforms the weak features and the strong features with 200 codebook size in [9], and is only inferior to the strong features with 400 codebook size. When a spatial pyramid is used, spatial PACT outperforms most existing methods. Spatial PACT is inferior to the strong SIFT features with 400 "intermediate concepts" in [19] and the SIFT features with 1200 pLSA topics in [7].

Note that length of the spatial PACT feature vector is only about 5% of the SP-pLSA feature vector length in [7]. By integrating more information into the spatial PACT feature vector, the recognition accuracy can be further improved. As shown in Fig. 12b, the Sobel gradient images contain information that emphasizes image discontinuities. By observing these Sobel

---

[7] The results reported here is slighted different from those in [40] in which the accuracy was computed as total number of correct predictions divided by total number of testing images. In Table IV the average accuracy in all categories are reported (i.e., average of diagonal entries in the confusion matrix).

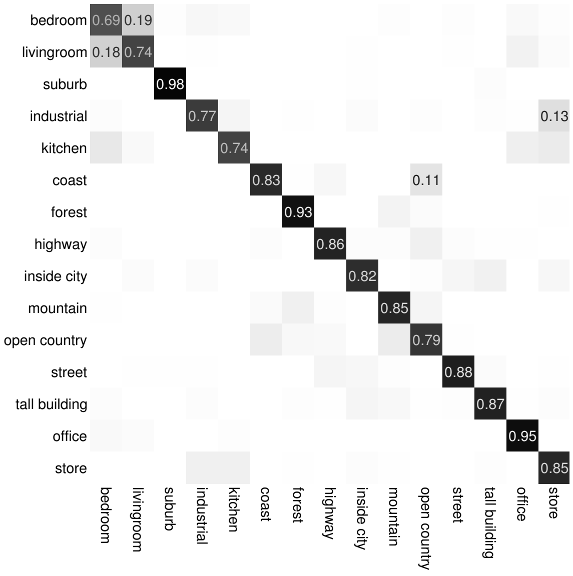


Fig. 13. Confusion matrix of the 15 class scene dataset. Only rates higher than 0.1 are shown in the figure.

gradient images, a person can gain confidence in predicting many scene category types. By generating spatial PACT feature vectors on Sobel images and concatenate them to the feature vectors generated from original images, spatial PACT (with Sobel) achieved a 84.96% accuracy, significantly higher than all other methods.

Confusion matrix from one run on this dataset ($L = 2$ spatial PACT) is shown in Fig. 13, where row and column names are true and predicted labels respectively. The biggest confusion happens between category pairs such as bedroom/living room, industrial/store, and coast/open country, which coincides well with the confusion distribution in [9].

More experiments were also carried out to compare our CENTRIST based descriptor with other descriptors, and to examine various aspects of the scene recognition problem.

Orientation Histogram. Orientation histogram [41] is a representation that uses histogram of quantities computed from 3×3 neighborhoods. We implemented this method with 40 bins. Combined with a level 2 spatial pyramid, Orientation Histogram achieves $76.78 \pm 0.90\%$ recognition rate, which is significantly worse than spatial PACT $(83.10 \pm 0.60\%)$.

Indoor-outdoor classification. We also distinguish indoor and outdoor scenes in this dataset. The *industrial* category contains both indoor and outdoor images, and is thus ignored. The remaining 14 categories are separated as 5 indoor categories and 9 outdoor categories. Using $L = 2$ and $L = 0$, spatial PACT successfully predicts labels for 98.02% and 95.31% of the

Fig. 14. Images from 8 different sports event categories.

images, respectively. These recognition rates are much higher than previous results on indoor-outdoor classification datasets (e.g. [23], [10]).

Linear classifiers. Linear SVM classifiers are also applied to the scene dataset. They achieve accuracy of 81.78% and 73.59%, using spatial PACT with $L = 2$ and $L = 0$, respectively. The implication of these results are two fold. First, the difference in performance of RBF kernels and linear kernels are quite small. In all the datasets we experimented with, the difference in recognition rates between these two kernel types are smaller than 2%. This observation suggests that images from the same category are compact in the spatial PACT descriptor space. Second, because of the fast testing speed of linear classifiers and small performance difference, linear SVM classifiers could be used to ensure real-time classification. A further observation is that linear SVM classifiers get 94.18% accuracy on indoor-outdoor classification, without using a spatial pyramid. In other words, CENTRIST could reliably distinguish the man-made indoor structures and the outdoor natural scenes.

Speed and classifier analysis. The time to extract CENTRIST is proportional to the input image size. However, large images can be down-sampled to ensure high speed. Our experiments observed only slight (usually $< 1\%$) performance drop. Also, spatial PACT is not sensitive to SVM parameters. $(C, \gamma) = (8, 2^{-7})$ is recommended for RBF kernels with probability output, and $C = 2^{-5}$ for linear SVM.

SVM vs. 1-NN. Finally, we want to point out that choosing the right classifier for a specific application is very important for scene recognition, too. If we use 1-NN for this scene recognition task, the level 2 spatial PACT features achieved a recognition rate of $63.84 \pm 1,21\%$, far below the SVM result (83.10%) with the same features.

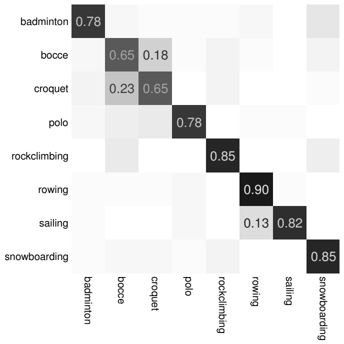


Fig. 15. Confusion matrix of the event dataset. Only rates higher than 0.1 are shown in the figure.

D. *The 8 class event dataset*

The event dataset [18] contains images of eight sports: badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snowboarding (see Fig. 14 for example images from each category). In [18], Li and Fei-Fei used this dataset in their attempt to classify these events by integrating scene and object categorizations (i.e. deduce *what* from *where* and *who*). We use this dataset for scene classification purpose only. That is, we classify events by classifying the scenes, and do not attempt to recognize objects or persons.

The images are high resolution ones (from 800x600 to thousands of pixels per dimension). The number of images in each category ranges from 137 to 250. Following [18], we use 70 images per class for training, and 60 for testing. The first 50 images in each category are used to compute the eigenvectors. We use RBF kernel SVM classifiers with level 2 pyramid spatial PACT features in this dataset.

Overall we achieve $78.50 \pm 0.99\%$ accuracy on this dataset. In [18], the scene only model achieved approximately 60% accuracy, which is significant lower than the spatial PACT result. When both scene and object categorization were used, the method in [18] had an accuracy of 73.4%, still inferior to our result. Note that this scene+object categorization used manual segmentation and object labels as additional inputs.

The scene only model of spatial PACT exhibits different behaviors than the scene+object model in [18], as shown in the confusion matrix in Fig. 15. The most confusing pairs of our method are bocce/croquet, and rowing/sailing. These results are intuitive because these two pairs of events share very similar scene or background. In [18], the most confusing pairs are bocce/croquet, polo/bocce, and snowboarding/badminton. The object categorization helped in distinguishing rowing and sailing. However, it seems that it also confused events that have distinct backgrounds, such as snowboarding and badminton.

Observations similar to those on the scene categorization dataset, are also present in the sports event dataset:

- Linear kernel SVM achieved almost indistinguishable accuracies as the RBF kernel SVM ($78.50 \pm 0.85\%$).
- A level 2 spatial PACT greatly improves a level 0 one, which has accuracy $64.67 \pm 1.71\%$.
- Adding the Sobel image improves the recognition accuracy to $80.54 \pm 0.64\%$.

E. Bag of Visual words with CENTRIST

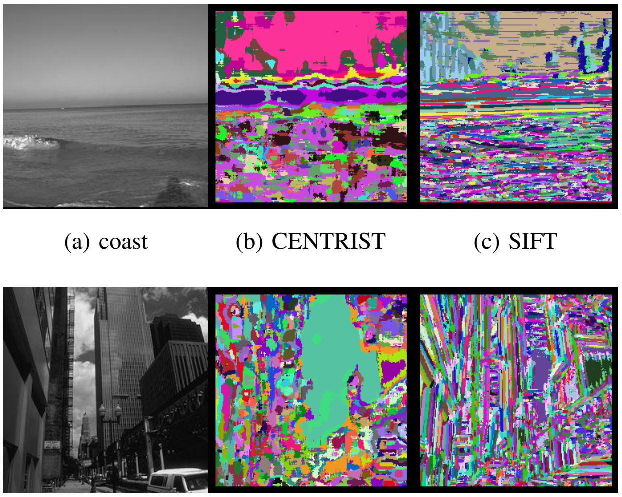
Since CENTRIST can be extracted for any rectangle image patches, we can also apply the Bag of Visual words framework with CENTRIST being the base visual descriptor. This is the second approach to use CENTRIST in this paper.[8]. Following [9], we use image patches of size 16 by 16, and sample over a grid with a spacing of 8 pixels. In every training/testing splitting, one fourth of the image patches sampled from the training set are used to generate a codebook which contain 200 codewords. Since CENTRIST is only 256 dimensions, PCA operations are not performed (i.e. CENTRIST is directly used for each 16 by 16 image patch). The k-means algorithm is used to cluster CENTRIST vectors into 200 codewords. For a level 2 spatial hierarchy, the final feature vector has a length of $200 \times (25 + 4 + 1) = 6200$. SVM classifiers with histogram intersection kernel are used.

On the 15 class scene recognition dataset, codebook of CENTRIST correctly recognize $80.73\pm 0.59\%$ of the testing images, which is similar to the result of codebook with 200 SIFT codewords in [9] ($81.1 \pm 0.3\%$), but inferior to the spatial PACT result ($83.10 \pm 0.60\%$). Similarly, on the 8 sports event dataset, codebook of CENTRIST achieved an accuracy of $75.21 \pm 1.06\%$, which is lower than the spatial PACT accuracy, but higher than those reported in [18].

Although the CENTRIST visual codebook's performance is not as good as spatial PACT, it provides a way to visualize the behavior of the CENTRIST descriptor, and consequently improve our understanding of CENTRIST. We build a visual codebook with 256 visual code words using the 15 class scene recognition dataset. Given an input image, an image patch with coordinates

---

[8] Source code is available at `http://www.cc.gatech.edu/cpl/projects/libHIK`



(a) coast (b) CENTRIST (c) SIFT

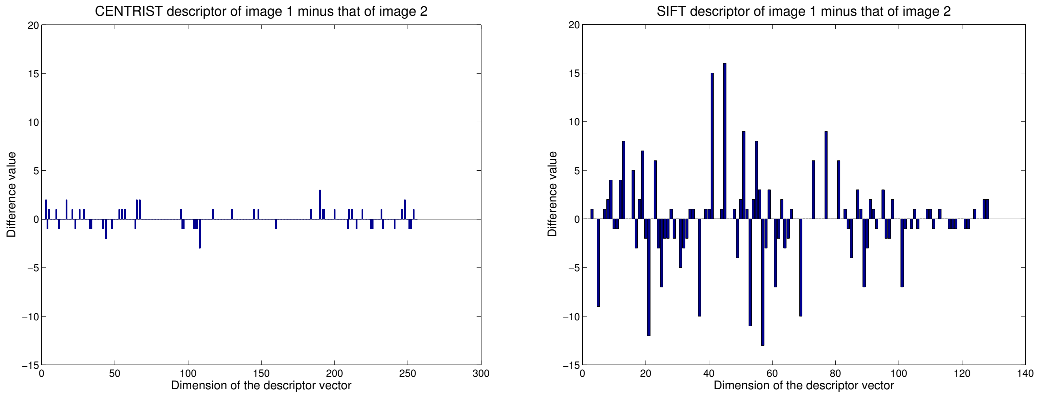
(d) tall building (e) CENTRIST (f) SIFT

Fig. 16. Visualization of images mapped to codewords. In each row, the first image is an input image, with the second and third being visualization for CENTRIST and SIFT codebooks, respectively. (This picture needs to be viewed in color.)

$[x-8, x+8) \times [y-8, y+8)$ can be mapped to a single integer by the following procedure. We first extract the CENTRIST descriptor from this window (whose size is 16 by 16). This CENTRIST vector is compared to all codewords, and the index of the nearest neighbor is the mapping result for pixel position (x, y). By choosing a random RGB tuple for each codeword index, a gray scale image can be transformed into a visualization of corresponding codeword indexes.

Fig. 16 are examples of the codebook visualization results for a coast and a tall building image. The SIFT code words tend to emphasize discontinuities in the images. Edges (especially straight lines) usually are mapped to the same codeword (i.e. displayed in the same color in the visualization). The visualization also suggests that SIFT pays more attention to detailed textural information, because the visualization is fragmented (connected component of the same color is small). Image patches with similar visual structure and semantics are mapped to different visual code words, e.g. the tall building in the right half of Fig. 16d.

Instead, CENTRIST visualizations tend to group image regions with similar visual structure into the same code word. The connected component in CENTRIST visualizations are larger than those in the SIFT visualizations. For example, the sky in the coast image share similar semantics and visual structures. This region is mostly mapped to the same color (i.e. same code word) using CENTRIST, which is desirable for the scene category recognition task. Instead, the SIFT



(a) CENTRIST (b) SIFT

Fig. 17. Difference vector of visual descriptors for Fig. 4b and Fig. 4c.

descriptor maps this region to different colors.

The different behaviors of CENTRIST and SIFT might be explained by the way local image measurements are accumulated. In CENTRIST, we only concern whether a center pixel's intensity is higher or lower than its neighbors. The magnitude of difference of pixel intensities is ignored. On the contrary, visual descriptors like SIFT and HOG accumulate orientation gradients of pixel intensities. The magnitude of pixel intensity differences has strong effect of histogram of orientation gradients. Thus, we conjecture that SIFT and HOG are more sensitive to smaller changes of visual contents than CENTRIST. In scene recognition we want our descriptors to be insensitive to small variations in images.

This conjecture is illustrated in Fig. 17. We examine two pictures that have similar structures: Image 1 as shown in Fig. 4b, and Image 2 as shown in Fig. 4c (in page 10). We compute the CENTRIST descriptors for $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ both images. The difference vector $\boldsymbol{h}_1 - \boldsymbol{h}_2$ is shown in Fig. 17a. We repeat the same procedure for the SIFT descriptor and the result is shown in Fig. 17b.

We require all feature vectors to have the same l_1 norm (i.e. all dimensions in a feature vector sum to the same constant). Thus the heights of bars in Fig. 17a and Fig. 17b are directly comparable. Although Image 1 and Image 2 have very similar image structures, their SIFT descriptors are quite different. A lot of bins have different values, i.e. having a non-zero value in the difference vector. The magnitude of the difference vector components are also big (some larger than 15). CENTRIST descriptors, however, are very similar in these two images. These differences are summarized in Table V, in which *sum of absolute difference* means $\sum_i |h_{1i} - h_{2i}|$.

TABLE V

DIFFERENCE OF VISUAL DESCRIPTORS FOR TWO SIMILAR IMAGES.

	Percent of changed cells	Sum of absolute difference
CENTRIST	21%	63
SIFT	67%	307

TABLE VI

COMPARING RECOGNITION ACCURACIES OF CENTRIST AND GIST IN SCENE RECOGNITION DATASETS.

Dataset	Environment	CENTRIST	Gist
8 class	outdoor	$85.65 \pm 0.73\%$	$82.60 \pm 0.86\%$
15 class	outdoor + indoor	$83.10 \pm 0.60\%$	$73.28 \pm 0.67\%$

F. Comparing CENTRIST, SIFT, and Gist

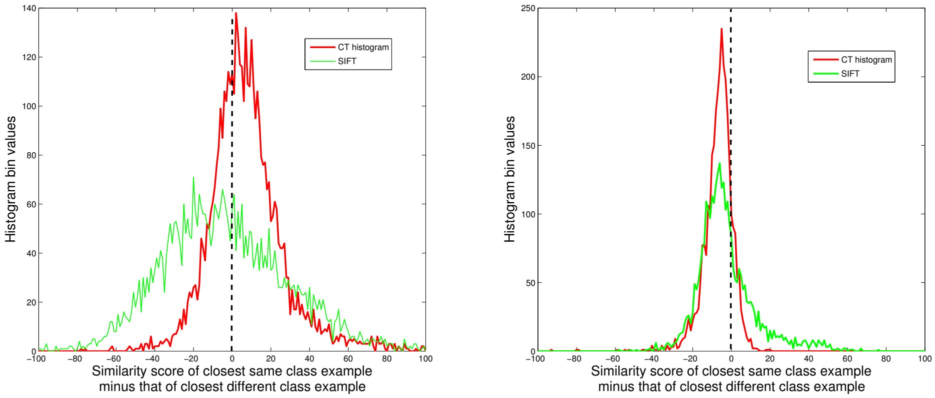
In this section we further compare the CENTRIST descriptor to the SIFT and Gist descriptors.

As mentioned in Sec. II, we observe that the perceptual properties Gist is modeling are mainly valid for outdoor environments. Our experiments on the 8 outdoor scene categories [10] and the 15 scene categories (which is a super set of the 8 category dataset) further corroborated this observations. Using the Gist descriptor[9] and SVM classifier, the recognition accuracy was $82.60 \pm 0.86\%$ on the 8 outdoor categories, which is worse than 85.65 ± 0.73, the accuracy using CENTRIST on this dataset. However, on the 15 class dataset which include several indoor categories, the accuracy using Gist dramatically dropped to $73.28 \pm 0.67\%$, which is significantly lower than CENTRIST's accuracy, $83.10 \pm 0.60\%$. Our conjecture is that the frequency domain features in the Gist descriptor is not discriminative enough to distinguish between the subtle differences between indoor categories, e.g. bedroom vs. living room. The same procedures and parameters are used in all experiments, except that CENTRIST and Gist are used in different experiments. Table VI summarizes these results.

On the contrary, SIFT is originally designed to have high discriminative power. Thus it may not be able to cope with the huge intra-class variation in scene images. For any two feature vectors, we can compute their Histogram Intersection Kernel (HIK) value [42] as a simple measure for the similarity between them. By observing the similarity distribution between- and within- categories, which are shown in Fig. 18 for both SIFT and CENTRIST, we can have an

---

[9] `http://people.csail.mit.edu/torralba/code/spatialenvelope/`



(a) scene recognition (b) Caltech 101

Fig. 18. Histogram comparing similarity values of best in-category nearest neighbor with best out-of-category nearest neighbor of an image. (This figure is best viewed in color.)

estimate of their capability in place and scene recognition.

For any image, we can find its nearest neighbor in the same category and the nearest neighbor in a different category. If the out-of-category nearest neighbor has a higher similarity value than the in-category nearest neighbor, the simple nearest neighbor classifier will make a wrong decision for this image. In Fig. 18 the x-axis shows the difference of these two similarity values. In other words, a value in the left hand side of 0 (the black line) means an error. For any given curve, if we find area of the part that is at the left hand side of the black dashed line, and divide it by area of the entire curve, we get the leave one out estimation of the classification error of a nearest neighbor rule. Thus Fig. 18 is an indication of the discriminative power of the descriptors. CENTRIST has a clear advantage in recognizing place and scene images (35.83% error, compared to 57.24% for SIFT), while SIFT is suitable for object recognition (67.39% error, compared to 83.80% for CENTRIST).

V. CONCLUSIONS

In this paper we propose CENTRIST, CENsus TRansform hISTogram, as a visual descriptor for recognizing places and scene categories. We first show that place and scene recognition pose different requirement for a visual descriptor, especially for such tasks in indoor environments. Thus we need a visual descriptor that is different from commonly used ones (e.g. SIFT in object recognition). We analyze these tasks and show that the descriptor needs to be holistic and generalizable. It also needs to acquire structural properties in the image while suppressing textural details, and contain rough geometrical information in the scene.

We then focus on understanding the properties of CENTRIST, and show how CENTRIST suits the place and scene recognition domain. CENTRIST is a holistic representation that captures the structural properties of an image. Through the strong constraints among neighbors Census Transform values, CENTRIST is able to capture the structural characteristic within a small image patch. In larger scale, spatial hierarchy of CENTRIST is used to catch rough geometrical information. CENTRIST also shows high generalizability, exhibiting similar visual descriptors for images with similar structures.

On four datasets including both place and scene category recognition tasks, CENTRIST achieves higher accuracies than previous state-of-the-art methods. Comparing with SIFT and Gist, CENTRIST not only exhibits superior performance. It has nearly no parameter to tune and is easy to implement. It also evaluates extremely fast. Implementation of methods proposed in this paper is publicly available at `http://www.cc.gatech.edu/~wujx/PACT/PACT.htm` and `http://www.cc.gatech.edu/cpl/projects/libHIK`.

In this paper we also analyzed several limitations of CENTRIST and there are research directions that may improve it. First, CENTRIST is not invariant to rotations. Although robot acquired images and scene images are usually upright, making it rotational invariant will enlarge its application area. Second, we want to recognize place categories in more realistic settings, i.e. learning the category concepts using images acquired without human effort in acquiring canonic views. Third, CENTRIST now only utilize the gray scale information in images. As shown in [43], different channels in the color space contain useful information for object and place recognition. The performance of CENTRIST should improve if color channels are incorporated appropriately.

ACKNOWLEDGMENTS

The authors would like to thank Henrik I. Christensen for fruitful discussions and Aaron Bobick for suggesting the Simulated Annealing optimization method. Special thanks to the many researchers who kindly published their datasets and software packages online.

REFERENCES

[1] B. Kuipers and P. Beeson, "Bootstrap learning for place recognition," in *AAAI Conference on Artificial Intelligence*, 2002, pp. 174–180.

[2] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust Monte Carlo localization for mobile robots," *Artificial Intelligence*, vol. 128, no. 1-2, pp. 99–141, 2001.

[3] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–108, 2006.

[4] S. Se, D. G. Lowe, and J. J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proc. IEEE Int'l Conf. Robotics and Automation*, 2001, pp. 2051–2058.

[5] I. Ulrich and I. R. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. IEEE Int'l Conf. Robotics and Automation*, 2006, pp. 1023–1029.

[6] H. Choset and K. Nagatani, "Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization," *IEEE Trans. on Robotics and Automation*, vol. 17, no. 2, pp. 125–137, 2001.

[7] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[8] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. II, 2005, pp. 524–531.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. II, 2006, pp. 2169–2178.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1575–1589, 2007.

[12] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen, "A discriminative approach to robust visual place recognition," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, 2006.

[13] J. Wolf, W. Burgard, and H. Burkhardt, "Robust vision-based localization for mobile robots using an image retrieval system based on invariant features," in *Proc. IEEE Int'l Conf. Robotics and Automation*, 2002, pp. 359–365.

[14] Z. Zivkovic and B. J. A. Kröse, "From sensors to human spatial concepts," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 357–358, 2007.

[15] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual Place Categorization: Problem, Dataset, and Algorithm," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, 2009.

[16] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[17] J. J. Kivinen, E. B. Sudderth, and M. I. Jordan, "Learning multiscale representaiton of natural scenes using dirichlet processes," in *The IEEE Conf. on Computer Vision*, 2007.

[18] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *The IEEE Conf. on Computer Vision*, 2007.

[19] J. Liu and M. Shah, "Scene modeling using Co-Clustering," in *The IEEE Conf. on Computer Vision*, 2007.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[22] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[23] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *CAIVD*, 1998, pp. 42–51.

[24] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *European Conf. Computer Vision*, 2008.

[25] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[26] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[27] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[28] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. II, 2003, pp. 264–271.

[29] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[30] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conf. Computer Vision*, vol. 2, 1994, pp. 151–158.

[31] D. Bhat and S. Nayar, "Ordinal measures for image correspondence," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 415–423, 1998.

[32] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[33] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *SIGGRAPH*, 1995, pp. 229–238.

[34] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training example: an incremental bayesian approach tested on 101 object categories," in *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[35] P. F. Felzenszwalb and J. D. Schwartz, "Hierarchical matching of deformable shapes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[36] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 286–299, 2007.

[37] O. J. O. Söderkvist, "Computer vision classification of leaves from swedish trees," Master's thesis, Linköping University, 2001.

[38] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The KTH-IDOL2 database," Kungliga Tekniska Hoegskolan, CVAP/CAS, Tech. Rep. CVAP304, October 2006.

[39] A. Pronobis and B. Caputo, "The KTH-INDECS database," Kungliga Tekniska Hoegskolan, CVAP, Tech. Rep. CVAP297, September 2005.

[40] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast asymmetric learning for cascade face detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 369–382, 2008.

[41] W. T. Freeman and M. Roth, "Orientation histogram for hand gesture recognition," in *FG workshop*, 1995, pp. 296–301.

[42] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[43] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for objects and scene recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.